L'analyse syntaxique de l'oral : problèmes et méthodes

Christophe Benzitoun, Estelle Campione, José Deulofeu, Sandrine Henry, Frédéric Sabio, Sandra Teston, André Valli, Jean Véronis

L'analyse syntaxique de l'oral: problèmes et méthode

Christophe Benzitoun, Estelle Campione, José Deulofeu
Sandrine Henry, Frédéric Sabio, Sandra Teston, André Valli, Jean Véronis

Equipe DELIC, Université de Provence
29, Av. Robert Schuman, 13621 Aix-en-Provence Cedex 1
Jean.Veronis@up.univ-mrs.fr

1. Introduction

Cette communication exposera les résultats de plusieurs mois de réflexion et d'expérimentation liées à la constitution d'un corpus oral de référence dans le cadre du projet d'évaluation des analyseurs syntaxiques Easy. La plupart des travaux sur l'analyse syntaxique automatique au cours des dernières décennies ont porté sur l'écrit, et l'on dispose de très peu de corpus oraux syntaxiquement annotés (à notre connaissance aucun pour le français). Or, de tels corpus seraient extrêmement intéressants, tant pour les études linguistiques, que pour l'évolution des technologies de la parole, dont les « modèles de langage » sont souvent mis au point à partir de textes écrits reflétant assez mal le langage parlé (par exemple le journal *Le Monde*). Dans une expérience que nous avons menée avec un système de reconnaissance automatique dans le domaine du renseignement ferroviaire, l'énoncé suivant :

> non non non non je veux pas de Pa- de de Paris gare d' Austerlitz

a par exemple été reconnu comme :

> Nancy dans nonante jours à Le Havre de Paris gare d' Austerlitz

Le système n'intègre pas la notion de répétition ou d'amorce de mot inachevé, et fournit donc, de façon erronée, une approximation lexicale au mieux de ses possibilités.

L'oral constitue un défi majeur pour l'analyse syntaxique, mais nous montrerons que les phénomènes que l'on y observe se retrouvent également pour beaucoup à l'écrit, même si c'est avec des fréquences moindres. Nous faisons donc l'hypothèse que l'étude de l'oral peut aussi amener quelque lumière dans les zones d'ombre, souvent négligées par commodité, de l'écrit.

2. Conventions de transcription

Le point de départ du corpus que nous mettons actuellement au point pour le projet Easy consiste en transcriptions d'oral, qui ont été effectuées par des experts avec un cycle de réécoute/validation extrêmement strict. Les conventions de transcription adoptées (DELIC,

2004) ne contiennent aucune ponctuation, suivant la tradition de Blanche-Benveniste & Jeanjean (1987), qui ont clairement montré que la ponctuation de l'écrit était parfaitement inadéquate à la transcription de l'oral (voir aussi Leech, McEnery & Wynne, 1997). Par contre, sont notés avec soin les pauses, les répétitions, les amorces (mots inachevés) et les *euh* d'hésitation. Dans le cas du corpus Easy, les pauses ont toutes été mesurées en millisecondes, et l'on a ajouté un marquage des mouvements intonatifs majeurs, ainsi que des allongements syllabiques d'hésitation (fonctionnellement équivalents au *euh*) (voir Campione, 2001, pour la méthodologie). Ces informations ne sont pas toujours ajoutées dans nos corpus, étant donné le coût considérable qu'elles engendrent, si l'on veut qu'elles soient fiables, mais elles paraissaient indispensables à l'action d'évaluation Easy : étant donné l'absence de ponctuation, il faut bien que les analyseurs aient à leur disposition les informations de structure présentes à l'oral.

La Figure 1 montre un exemple typique de transcription « brute », c'est-à-dire avant analyse syntaxique.

> on a un poste consacré aux objets sonores ↗ + donc là ça va de: {bruit : le locuteur tape sur un objet} ça hein des verres enfin tout un tas tout un tas de choses qui sont amplifiées retraitées ↗ par des effets rediffusés ↗ + et: un quatrième poste ↗ qui est surtout ce qui est: qui est: dédié aux: ce qu'on appelle les machines tournantes ↘ donc tout c- tout ce qui est enregistrements analogiques numériques les bandes les disques durs les machins ↗ enfin tout ce qui peut: tout ce qui peut s'enregistrer ↗ + donc qui consiste euh qui consiste à les: + à travailler sur des supports en différentes vitesses ↗ à manipuler à la main des bandes magnétiques analogiques ↗ pour jouer sur des vitesses obtenir des formes des choses comme ça ↗ + qui consiste aussi à les: à les repiquer à la volée des: des petits morceaux de ce que sont en train de faire les autres ↗ pour les retraiter différemment sur des boucles ou des choses comme ça ↗ pour ré-impulser ↗ + pour ré-impulser de la matière sur laquelle les autres vont réagir ↗ et puis et vice versa l'interaction elle est là ↘

Figure 1. Un exemple de transcription

La Figure 2 donne quelques-unes des notations utilisées dans les transcriptions.

+	pause	↗	intonation montante
xx	amorce (mot commençant par *xx*)	↘	intonation descendante
-			
xx:	allongement syllabique d'hésitation	→	intonation plate
{ }	événement non linguistique		

Figure 2. Notations

3. Phénomènes de production

La première particularité qui saute aux yeux dans les transcriptions d'oral est la présence importante de phénomènes liés à la production de l'oral (hésitations, amorces, répétitions, constructions interrompues, anacoluthes, etc.), même chez les « professionnels de la parole » (journalistes, hommes politiques, etc.). Ces phénomènes sont souvent appelés « disfluences », mais ce terme semble évoquer une anormalité, alors que les phénomènes concernés font partie des modes de production tout à fait normaux de l'oral. En fait, parler de disfluences revient à poser l'écrit normé comme modèle de la « fluence » normale, ce que nous souhaitons éviter. La Figure 3 met en évidence en caractères gras les phénomènes de production dans l'exemple de la Figure 1.

on a un poste consacré aux objets sonores ↗ + donc là ça va de: {bruit : le locuteur tape sur un objet} ça hein des verres enfin **tout un tas** tout un tas de choses qui sont amplifiées retraitées ↗ par des effets rediffusés ↗ + et: un quatrième poste ↗ **qui est surtout ce qui est:** qui est: dédié **aux:** ce qu'on appelle les machines tournantes ↘ donc **tout c-** tout ce qui est enregistrements analogiques numériques les bandes les disques durs les machins ↗ enfin **tout ce qui peut:** tout ce qui peut s'enregistrer ↗ + donc **qui consiste euh** qui consiste **à les:** + à travailler sur des supports en différentes vitesses ↗ à manipuler à la main des bandes magnétiques analogiques ↗ pour jouer sur des vitesses obtenir des formes des choses comme ça ↗ + qui consiste aussi **à les:** à les repiquer à la volée **des:** des petits morceaux de ce que sont en train de faire les autres ↗ pour les retraiter différemment sur des boucles ou des choses comme ça ↗ **pour ré-impulser** ↗ + pour ré-impulser de la matière sur laquelle les autres vont réagir ↗ **et puis** et vice versa l'interaction elle est là ↘

Figure 3. Phénomènes de production

Les constructions interrompues ont été soulignées. Celles-ci sont parfois « réparées » (*qui consiste à **les :** à travailler*), parfois pas (*qui est: dédié **aux:** ce qu'on appelle les machines tournantes*). Dans ce dernier cas, la construction résultante semble déviante — mais la notion de déviance demande à être analysée plus en détail, car elle requiert la référence à une norme, et il serait naïf de prendre pour norme l'écrit standard. Si des structures sont régulières à l'oral, elle doivent être prises en compte en tant que telles.

Notre équipe a lancé un programme systématique d'étude des phénomènes de production à l'aide des grands corpus informatisés dont elle dispose : étude des répétitions (Henry, 2002), des hésitations (Campione & Véronis, 2004), et des amorces (Pallaud, 2002) et des régularités commencent à émerger : on observe que les répétitions frappent principalement les mots grammaticaux introducteurs de syntagmes (prépositions, articles), et que des interactions très fortes lient les différents phénomènes de production (Henry & Pallaud, 2003 ; Campione & Véronis, 2004 ; Henry, Campione & Véronis, 2004), qui permettent d'envisager la mise au point de modèles prédictifs pour leur identification automatique.

Dans le cadre de la mise au point du corpus Easy, le marquage se fait de façon manuelle, et les phénomènes de production sont balisés, avec leur type, et éventuellement la « réparation » manquante (*aux* → *à* dans l'exemple précédent).

4. Segmentation

La plupart des travaux de traitement automatique des langues considèrent la phrase comme une unité naturelle. La pertinence linguistique de cette notion fait pourtant pour le moins l'objet d'un débat (voir par exemple : Berrendonner, 2002 ; Blanche-Benveniste, 2002 ; Kleiber, 2003). On note d'ailleurs que même à l'écrit, phrases et unités linguistiques sont parfois non concordantes. Dans l'exemple de la Figure 4, tiré du journal *Libération*, la première phrase contient deux unités linguistiques indépendantes, marquées par || (on peut bien sûr les considérer comme dépendantes par « juxtaposition », mais n'est-ce pas une tentative désespérée de sauver l'habitude ?). Par contre, là où le texte marque une rupture entre deux phrases (signe ═), il n'y a qu'une seule unité linguistique en jeu (*encore plus aujourd'hui qu'hier* est le complément du verbe *aime*).

Soyons direct, ‖ après l'avoir fréquenté depuis des années, après l'avoir écouté pendant des heures au long de monologues sans fin, on aime Claude Got. ═ Et peut-être encore plus aujourd'hui qu'hier, alors qu'on l'accuse de vouloir imposer une « société sanitaire », sans plaisirs ni risques, ennuyeuse à mourir. [Libération, 9 mars 2004]

Figure 4. Non concordance des unités linguistiques et des phrases

De nombreux autres cas de divergence existent (listes, etc. : cf. Gala, 2003). La phrase est au mieux une approximation graphique, qui résulte d'un compromis entre structure syntaxique, intonation et mise en page. A l'oral, les majuscules et les points n'existent pas, et la notion de phrase y est encore moins opératoire.

Nous avons choisi de suivre une approche analogue à celle de Blanche-Benveniste (2002), qui part des constructions verbales et non de la phrase pour définir des unités de segmentation. La référence à la construction verbale est très adaptée au monologue narratif ou explicatif ; dans les conversations, les structures de références sont souvent non verbales (*non, oui, d'accord, la semaine prochaine,* etc.). Nous définissons donc des « unités maximales » (UM), qui sont des constructions verbales, nominales, adjectivales ou adverbiales regroupant un élément tête ainsi que tous les éléments qui sont dans sa dépendance. Nous analyserons ainsi l'exemple de la Figure 4 comme étant composé de deux UM (séparées par le signe ‖). L'exemple de la Figure 1 est, quant à lui, composé de trois UM :

① on a un poste consacré aux objets sonores ↗ +

② *donc là ça va de: ça hein des verres enfin* ***tout un tas*** *tout un tas de choses qui sont amplifiées retraitées ↗ par des effets rediffusés ↗ +*

et: un quatrième poste ↗ **qui est surtout** **ce qui est:** qui est: dédié **aux:** ce qu'on appelle les machines tournantes ↘ donc **tout c-** tout ce qui est enregistrements analogiques numériques les bandes les disques durs les machins ↗ enfin **tout ce qui peut:** tout ce qui peut s'enregistrer ↗ + donc **qui consiste euh** qui consiste **à les:** + à travailler sur des supports en différentes vitesses ↗ à manipuler à la main des bandes magnétiques analogiques ↗ pour jouer sur des vitesses obtenir des formes des choses comme ça ↗ + qui consiste aussi **à les:** à les repiquer à la volée **des:** des petits morceaux de ce que sont en train de faire les autres ↗ pour les retraiter différemment sur des boucles ou des choses comme ça ↗ **pour ré-impulser** ↗ + pour ré-impulser de la matière sur laquelle les autres vont réagir ↗ **et puis** et vice versa

③ l'interaction elle est là ↘

Figure 5. Découpage en unités maximales

On note que la totalité de l'UM ①, malgré sa taille, est sous la dépendance du verbe *avoir* de *on a*. L'UM ② est enchâssée en incise dans l'UM ①. On trouve également des parenthétiques à l'écrit (marquées entre tirets, entre parenthèses, ou simplement entre virgules), mais elles sont très fréquentes à l'oral. Elles se caractérisent généralement par une prosodie marquée (intonation plate basse, débit élevé, etc., cf. Campione, 2001).

5. Listes

L'organisation sous forme de « listes » (Blanche-Benveniste, 1990) est omniprésente à l'oral. Par liste, nous entendons un ensemble de constituants qui occupent la même place syntaxique dans une UM (Figure 6). Le phénomène existe à l'écrit (et il est même assez répandu dans certains styles). Néanmoins, sa fréquence est frappante à l'oral, à tel point qu'on ne peut simplement se contenter d'y voir un détail négligeable.

on a
un poste consacré ...
<u>et</u> un quatrième poste

tout un tas de choses qui sont
amplifiées
retraitées par des effets rediffusés

et un quatrième poste
qui est est dédié ...
<u>donc</u> qui consiste à travailler ...
qui consiste aussi à les repiquer ...

Figure 6. Listes

Les listes ont des valeurs diverses, et il est souvent très hasardeux d'en faire une analyse en termes de coordination : elles peuvent être aussi des rattrapages de production (le locuteur répare), des ajouts d'information (le locuteur précise), des jeux de modalité (le locuteur oppose ou compare), des conclusions (le locuteur résume), etc. Des « joncteurs » viennent souvent introduire certains termes des listes (*enfin, pas, mais, et, je veux dire, donc, sinon,* etc.) :

en fonction
de +
de la la possibilité +
<u>pas</u> des capacités +
<u>mais</u> des possibilités +

La coordination apparaît comme un cas particulier de liste.

Les listes peuvent s'imbriquer à plusieurs niveaux, de façon tout à fait surprenante. La Figure 7 montre que le locuteur utilise à deux reprises une imbrication de liste à cinq niveaux, tout en incluant une parenthétique ! Nous avons dans nos corpus des exemples d'imbrications encore plus complexes, et il est frappant de constater que les locuteurs perdent rarement le fil de ces architectures complexes.

Nous considérons les listes non comme un épiphénomène sans importance, mais comme un principe structurant de même nature que les relations de dépendance. Les arbres de dépendances se développent classiquement dans un plan. Les listes développent les arbres dans une troisième dimension par un processus de « marcottage » qui engendre de nouveaux arbres de dépendances à partir de certains noeuds.

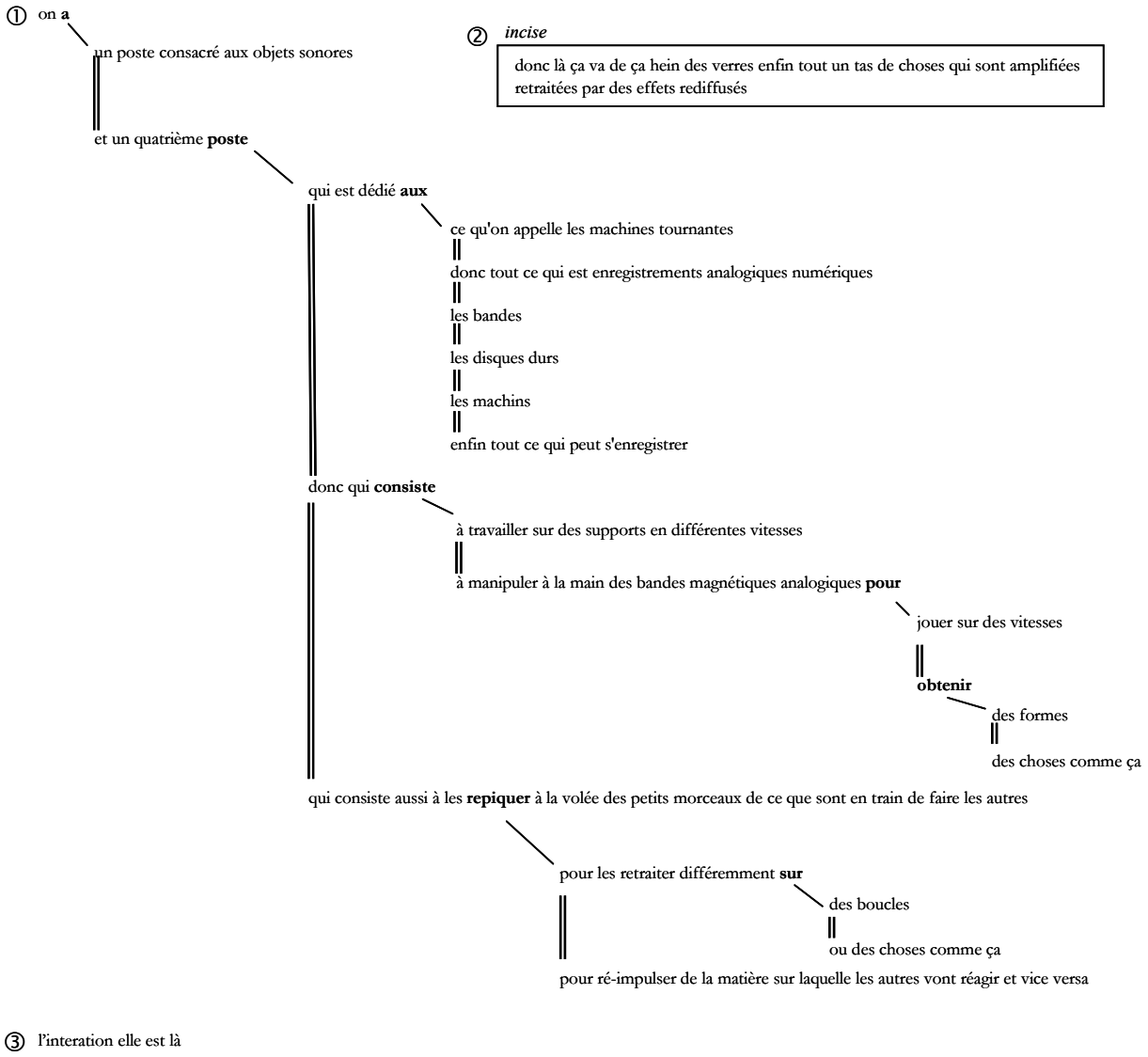

Figure 7. Imbrication de listes

6. Inserts

Une autre caractéristique de l'oral est l'omniprésence de mots ou locutions tels que *hein, bon*, *ben, quoi, tu vois, tu sais,* que l'on désigne par des termes aussi divers que « particules discursives », « marqueurs discursifs », etc. Nous adoptons la terminologie de Biber et al. (1999 : 93-94 et 1082-1083), qui parlent d' « inserts ». Les inserts sont définis comme étant des mots qui n'entrent dans une construction syntaxique avec aucun autre élément, tout en étant attachés prosodiquement à l'UM dans laquelle ils prennent place :

> oui mais c'est pas définitif **hein** madame votre question
> le problème c'est que **bon** ça ça stoppe + quand on écarte les bras
> en pleine campagne trouver des choses à faire **ben** c'est devenir agriculteur ou éleveur **quoi**
> c'était un système d' habitude qui s' était instauré une routine **tu vois** + et j' étais pas amoureux
> on n'avait pas pas amené au magasin le petit bout de tapisserie **tu sais** pour voir la couleur

Les inserts sont problématiques pour l'analyse automatique car ils ont souvent des homographes. Par exemple, *quoi* peut être également pronom, *bon* peut être adjectif, adverbe

ou nom, *tu vois* et *tu sais* peuvent être recteurs d'une véritable construction. La distribution s'inverse entre écrit et oral (Teston & Véronis, 2004). La Figure 8 montre la proportion d'usages du mot *bon* comme insert dans quatre corpus de 440 000 mots chacun de genres différents (oral, forum, littérature, presse). On voit que cet usage n'est pas totalement exclu à l'écrit (en particulier sur les forums : les « nouvelles formes de communication écrite » ont tendance à reprendre certaines caractéristiques du langage parlé), mais qu'il est prépondérant à l'oral.

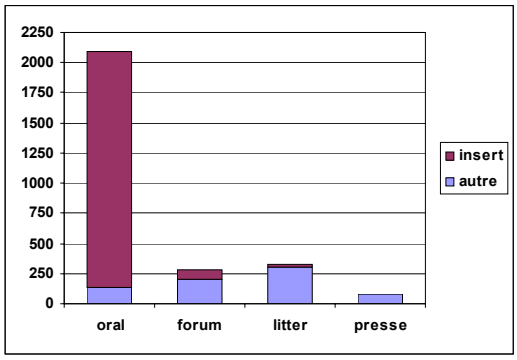

Figure 8. *Bon :* proportion d'usages comme insert

Dans une perspective de constitution de corpus analysé, les inserts doivent être repérés et isolés de l'arbre de dépendances.

7. Conclusion

De prime abord, les transcriptions d'oral apparaissent comme d'une complexité décourageante pour la constitution de corpus de référence syntaxiquement annotés. Cette communication essaie toutefois de montrer qu'en procédant avec méthode, cette complexité disparaît. Nous procédons en trois étapes :

- identification et marquage des phénomènes de production ;
- identification des listes et organisation de l'énoncé en arbres « marcottés » ;
- identification et isolement des inserts.

A l'issue de ces trois étapes, l'objet résultant est beaucoup plus familier (Figure 7), et se prête au marquage des relations de dépendances d'une façon parfaitement analogue à celle de l'écrit. On retrouve ici l'une des positions méthodologiques de notre équipe : la syntaxe de l'oral ne diffère en rien de celle de l'écrit, sauf, sans doute en termes de proportions. Ainsi on trouvera à l'oral beaucoup de clivées (*c'est le coiffeur qui est content*) et pseudo-clivées (*ce qui l'intéresse c'est le pognon*), de doubles marquages (*je vous en ai pas parlé du quartier d'isolement*), de fausses subordonnées (*c'est assez artificiel de les regrouper l'un à côté de l'autre + parce qu'il faut savoir que les coraux s'attaquent entre eux*), etc. Toutefois, ces constructions ne sont pas du tout limitées à l'oral, et l'oral nous obligera peut-être, à cause de leur fréquence, à ne pas les négliger en leur plaquant des analyses fausses (et non « théoriquement neutres »), dont on se satisfait parfois, par commodité, pour l'écrit.

Références

Berrendonner, A. (2002). Les deux syntaxes. *Verbum.* XXIV(1-2):23-35.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. (1999). *Longman Grammar of Spoken and Written English.* London : Pearson ESL.

Blanche-Benveniste, C. (Ed.) (1990). *Le français parlé - Etudes grammaticales*, Paris : C.N.R.S.

Blanche-Benveniste, C. (2002). Phrase et construction verbale. *Verbum.* XXIV(1-2):7-22.

Blanche-Benveniste, C., & Jeanjean, C. (1987). *Le français parlé. Edition et transcription.* Paris : Didier-Erudition.

Campione, E. (2001). *Etiquetage prosodique semi-automatique de corpus oraux : algorithmes et méthodologie. Thèse de doctorat.* Aix-en-Provence: Université de Provence.

Campione, E., & Véronis, J. (2004). Pauses et hésitations en français spontané. *Actes des XXVè Journées d'Etude sur la Parole (JEP'2004)* (sous presse). Fès (Maroc).

Gala, N. (2003). Un modèle d'analyseur syntaxique robuste fondé sur la modularité et la lexicalisation de ses grammaires. *Thèse de doctorat en informatique.* Université de Paris-Sud.

Henry, S. (2002). Etude des répétitions en français parlé spontané pour les technologies de la parole, *Actes de la 6ème Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL'02)* (pp. 467-476). Nancy (France) : ATALA.

Henry, S., Campione, E., & Véronis, J. (2004) Répétitions et pauses (silencieuses et remplie) en français spontané. *Actes des XXVè Journées d'Etude sur la Parole (JEP'2004)* (sous presse). Fès (Maroc).

Henry, S., & Pallaud, B. (2003). Word fragments and repeats in spontaneous spoken French, *Proceedings of DiSS'03, Disfluency in Spontaneous Speech Workshop* (pp. 77-80). Göteborg (Sweden).

Kleiber, G. (2003). Faut-il dire adieu à la phrase ? *L'information grammaticale, 98*:17-22.

Leech, G., McEnery, A., & Wynne, M. (1997) Further levels of annotation. In Garside, R., Leech, G., & McEnery, A. (Eds) *Corpus Annotation* (pp. 85-101). London: Longman.

Pallaud B. (2002). Les amorces de mots comme faits autonymiques en langage oral, *Recherches sur le français parlé, 17,* 79-102.

Pallaud, B., & Henry, S. (2004). Amorces de mots et répétitions: des hésitations plus que des erreurs en français parlé spontané, *Actes des 7èmes Journées internationales d'Analyse statistique des Données Textuelles (JADT'2004)* (pp. 848-858). Louvain-la-Neuve (Belgique).

Teston, S., & Véronis, J. (2004). Recherche de critères formels pour l'identification automatique des particules discursives. *Modéliser et décrire l'organisation discursive à l'heure du document numérique.* Journée ATALA, La Rochelle, 22 juin 2004 (Semaine du Document Numérique).